The Impossibility of the Happiness Pill

Bentham, Mill, and Isaiah Berlin on Determinism and Liberal Neutrality

Gianfranco Pellegrino
Bentham Project (UCL)
uczwgpe@ucl.ac.uk

(Abstract. Isaiah Berlin once claimed that Bentham and James Mill might have considered the use of "techniques of subliminal suggestion or other means of conditioning human beings." In this paper I argue that Berlin's provocative suggestion is quite misplaced since Bentham wouldn't have given such medical treatment, even if it were possible, as many passages of his most known texts make clear. In the context of his criticism of moral sense theories, Bentham claimed that there is no stable connection between motives and actions, because the same sort of motives could lead to different actions, depending on the sensibilities of the agent. This view of motives and their connection with actions makes empirically impossible the nightmare of an illiberal mass treatment with medicines: if there are no stable connections between motives and actions, it is impossible for any medicine to have uniform effects over persons. Bentham's skeptical arguments provide strong foundations to a liberal view of the limits and the legitimacy of government intervention. Bentham endorsed the well-known 'best judge argument' in favor of non-interference, providing a distinctive and novel foundation to it.

1. Manipulative Utilitarianism: Berlin on Bentham and Mill

According to a common view, the following argument hold true:

- **The manipulation argument** For utilitarianism, an action is right if and only if it is conducive to the greatest happiness of the greatest number. Means to this end matter only for their consequences, and in particular for their impact on the final amount of happiness produced. No means are intrinsically wrong. For instance, manipulative techniques aimed at curbing anti-felicitic behavior are not intrinsically impermissible. Hence, artificial, paternalistic, hidden or unconscious barriers to anti-felicitic behavior are not intrinsically impermissible. All that matters is that, in doing so, more happiness than unhappiness is produced overall.

Relying on the argument above, utilitarianism is criticized for being necessarily anti-liberal. The objection rests on the assumption that manipulation is intrinsically wrong, because it cannot be reconciled with freedom or autonomy, or with treating individuals as ends, rather than merely as means.1 Basically, being autonomous or free amounts to acting, and to deciding how to act, without being manipulated by others—and any manipulation of this sort diminishes the freedom of the manipulated subject.2 Permitting manipulation of some individuals for the sake of the overall happiness, utilitarianism denies that freedom and autonomy have intrinsic value. Such denial puts utilitarian views in the field of non-liberal or anti-liberal political theories.

1. See Berlin (2002c, 17).
2. See Berlin (2002c, 18).
This charge has been pressed by Isaiah Berlin in his paper on John Stuart Mill. The main theme of Berlin's paper is the contrast between the conception of happiness and human nature that Mill advocated in *On Liberty* (1859), and the anthropological and epistemological views defended by Bentham and James Mill. Berlin claims that on Mill's account autonomy and freedom of choice are necessary ingredients of happiness. On Berlin's interpretation, Mill derived this view of happiness from anthropological and anthropological premises. According to Mill, human nature and truth are far from being definitive, fixed once for all, or ascertainable in a conclusive way. Consequently, openness to eccentric choices and opinions, along with the capacity to assess different theories and to experience different lifestyles, as well as to autonomously choose them, are necessary components of the individual pursuit of happiness: nobody can fix in advance what will make someone happy. Accordingly, individuals should be let free to individually search for their recipe for a happy life. Moreover, Mill argued that the capacity to choose autonomously and to self-determine oneself is the only permanent feature and the distinguishing mark of human nature. Accordingly, any limitation of freedom and autonomy denies humanity and debases the person who is so limited and constrained, literally degrading human nature. Berlin points out two things: first, this kind of anthropological and epistemological fallibilism is a major departure from the utilitarian science of nature that Mill inherited from his father and Bentham; second, John Stuart Mill's plea for human liberty, individuality and autonomy rests on his rejection of the strict determinism endorsed by Bentham and James Mill.

The utilitarian framework endorsed by Bentham, as Berlin sees it, mainly relied on the idea that human nature has fixed features, albeit with some room for gradual change, and on a deterministic view of action and will. Bentham and James Mill adopted "the pseudo-scientific model, inherited from the classical world and the age of reason, of a determined human nature, endowed at all times, everywhere, with the same unaltering needs, emotions, motives, responding differently only to difference of situations and stimulus, or evolving according to some unaltering pattern." Like many Enlightenment thinkers, Bentham and James Mill claimed that education and legislation are the best means to harmonize the opposite interests of individuals and to drive people to bring about the greatest happiness of the greatest number. However, utilitarian education and legislation are successful only because of the possibility to employ certain fixed characteristics of human

---

3 See Berlin (2002b).
4 See Berlin (2002b, 221).
5 See Berlin (2002b, 222).
6 Berlin often employed a similar fallibilist view as a premise for his value pluralism and his liberalism: the faith in a single truth, independent of discussion and debate, often paved the way to totalitarianism; accordingly, liberalism requires skepticism and its ontological counterpart, value pluralism; see, for instance, Berlin (2002d, 345-6). On the other hand, value pluralism makes necessary (negative) freedom, as a free choice among competing values is needed: see Berlin (2002f, 212-7) and Harris (2002, 354).
7 See Berlin (2002b, 221-3, 227, 230-9, 249) and (2002f, 214).
8 Berlin (2002b, 249).
beings (for instance, their desire for pleasure and aversion to pain or the associative relations between certain ideas or feelings in the mind) to drive people towards happiness-conducive behavior. No means able to achieve this end is intrinsically wrong. For Bentham and James Mill, there are no differences between education and manipulation, between legislation and coercion. To put it better, the only differences are due to contingent obstacles that make manipulation or coercion too difficult, or unable to maximize happiness. Berlin stressed that, because of this, Bentham and James Mill were no genuine liberals, and their declarations in favor of the best-judge principle (according to which any individual is the best judge of her own interest; therefore, state paternalist intervention is always counter-productive) are only the outcome of psychological generalizations, whereas for John Stuart Mill individualism and the defense of individual freedom is a normative ideal.

Berlin's interpretation of pre-Millian utilitarianism is apparent in the following passage:

James Mill and Bentham had wanted literally nothing but pleasure, obtained by whatever means were the most effective. If someone had offered them a medicine which could scientifically be shown to put those who took it into a state of permanent contentment, their premises would have bound them to accept this as the panacea for all that they thought evil. Provided that the largest number of men receives lasting happiness, or even freedom from pain, it should not matter how this is achieved. Bentham and James Mill believed in education and legislation as the roads to happiness. But if a shorter way had been discovered, in the form of pills to swallow, techniques of subliminal suggestion or other means of conditioning human beings [...] then, being men of fanatical consistency, they might well have accepted this as a better, because more effective and perhaps less costly, alternative than the means that they had advocated. John Stuart Mill, as he made plain both by his life and by his writings, would have rejected with both hands any such solution. He would have condemned it as degrading the nature of man.

On Berlin's account, Bentham and James Mill reached anti-liberal conclusions because of their deterministic view of human agency. Manipulation is possible if human beings are understood as immersed in the network of natural causes and manipulative techniques provide means to determine human behaviour. An indeterministic or libertarian view of agency makes manipulation impossible, and it is clearly a necessary premise of liberalism.

Here, I am not interested in the general issue of determinism vs. indeterminism or libertarianism, or in giving a picture of the connection between metaphysical versions of libertarianism and liberalism. Rather, I shall focus on a historical and interpretive issue of Bentham's view, setting aside James Mill's position. So my

---

9 See Berlin (2002b, 220, 222).
10 See Berlin (2002b, 223).
11 Berlin (2002b, 222).
12 Berlin's view of the relevance of indeterminism or libertarianism is in Berlin (2002c, 4-30) and (2002e); see also Harris (2002, 349, 352-3, 354). On Berlin's interpretation of Mill, see Bellamy (1992), Harris (2002, 361-2), Wolheim (1979).
question is: Did Bentham endorse the strict deterministic view of human agency that Berlin ascribed to him? My answer will be in the negative. Bentham did not regard human agency as strictly determined. Indeed, his view of the connection between motives and action tended to be indeterministic. In the next three sections, I shall consider Bentham's indeterministic view of the connection between the will, motives and action. In particular, I shall show that Bentham rejected determinism, both in general and in his theory of action (§§ 2-3). Then, in §§ 4-5, I shall take on how Bentham's indeterministic view of the connection between motives and intentions grounded his liberalism.

2. Bentham's indeterminism I: the relation of cause and effect as a fiction

Determinism can come in different versions, and can be interpreted in many ways. For the purposes of this paper, I'll assume that determinism consists of two (very general) claims:

- **causal closure**: causal chains are closed, that is, there are no uncaused effects, or causes without effects;
- **causal determination**: kinds of causes have *uniquely* determined effects, i.e. necessarily each kind of cause produces (or contributes to the production of) its typical kind of effect.

Causal determination can be viewed as a restriction on causal closure. Causal closure entails that each token-cause should have some token-effect, even though there is no necessity that the same kind of causes produces the same kind of effects. Causal determination drops this permission, by dictating that each given kind of cause should be the *unique* cause of a given kind of effect, its *typical* effect. Causal determination presupposes causal closure, but the latter does not entail the former. It is logically possible to defend causal closure without endorsing causal determination. Indeed, it is my contention here that Bentham assumed causal closure, but rejected causal determination.

A deterministic view of action naturally follows from the two claims above; if there are no uncaused events, and any cause has its typical effect, then it cannot be the case that human actions are uncaused, or that they are caused by the human will, self or agency, where the will, self or agency are in their turn uncaused springs of action. Human actions should lie within complex causal chains, being caused by something else, and so on. Accordingly, if one can know the relevant causal chain, human actions are predictable and determined. Understood as the possibility to act otherwise, by breaking the causal chain that would have led to a determined action, freedom is impossible.

As stated above, causal closure is a very general claim, and needs specification. When specified, it can bring about alternative specific views. For instance, causal closure can be viewed as a premise of

- **causal physicalism**: there are no non-physical causes.

However, there is no necessary entailment from causal closure to causal physicalism. For causal closure can also be a premise of
causal dualism: there are non-physical, for instance mental, and material, or physical, causes, and causation can be transmitted across the two realms of immateriality and physicality (for instance, mental causes can have physical effects, and vice versa).\textsuperscript{13}

Causal dualism is different from:

- causal separatism: there are immaterial (for instance mental) and material (or physical) causes, but causation cannot be transmitted across the two realms of immateriality and physicality (for instance, mental causes cannot have physical effects, and vice versa).

Causal separatism entails a weak form of indeterminism, because it allows that some causes may be incapable of producing relevant effects—for instance, mental causes may be incapable of producing physical effects, and physical causes may be unable to produce mental effects. If so, it might be admitted that there are at least some causes devoid of effects, or better some causes without effects in each possible ontological realm. Causal separatism can be reconciled with determinism if a different version of causal closure is assumed, namely

- ontologically embedded causal closure: in each ontological realm, causal chains are closed, namely there are no uncaused effects in given ontological context, i.e. causes without effects belonging to that realm.\textsuperscript{14}

Bentham did not explicitly stated any of the principles above. However, in some of the remarks he made on the notion of 'cause' he seemed to reject causal closure. In particular, his view was that it is epistemologically impossible to endorse such a principle, because:

i. the notion of 'cause' is a mere linguistic convention, i.e. a fictitious entity;\textsuperscript{15}

ii. very often, the notion is inapplicable. For instance, when universal gravitation is considered, there is no point in taking each body attracting the other as a cause, and if one considers infinite motions in the universe, there is no point in isolating some event as an effect:

Each body attracts towards it all the rest: and were it to have place singly, the attraction thus exercised might be considered as if it operated in the character of a cause. But each body is attracted by every other: and were it to have place singly, the attraction thus suffered might be considered in the character of an effect. But in fact the two words are but two different names for one and the same object—for one and the same motion or tendency to motion. [...] No such character as that of

\textsuperscript{13} From the logical point of view, the realm of immaterial, or non-physical, entities should not be limited to mental entities. It might be admitted that the non-physical realm contains immaterial, but not mental, entities. Moreover, to establish whether mental events are physical or not is one of the issues at stake in this debate. However, one of the most discussed, and defended, position is that the realm of immaterial entities overlaps completely with the mental sphere.

\textsuperscript{14} This account of causal closure is in part inspired, despite differences in formulations, by Lowe (2000).

\textsuperscript{15} Bentham (1997, 102).
agent—no such character as that of patient—belongs separately to any one: each
one is agent and patient at the same time: no one exhibits more of agency, no one
more of patiency, than any other.

Suppose that, all these several bodies having been created out of nothing at one
and the same instant, each with the same quantity of matter and thence with the
same attractive power that appears to belong to it at present, an impulse in a
certain rectilinear direction were to be given to each of them at the same time: on
this hypothesis it has been rendered, it is said, matter of demonstration that the sort
of intermediate motions which would be the result would be exactly those which
these same bodies are found by observation to exemplify.

Here then we should have a beginning: but even here we should not have an end.
In the beginning, at a determinate point of time, we should have a motion
operating in the character of a cause: but at no determinate point of time to the
exclusion of any other should we have either a motion or a new order of things
resulting from it in the character of an effect.16

In this discussion of the notion of 'cause', Bentham also claimed that causal
determination is involved in the common meaning of causal concepts, and that this
view is wrong:

In the use commonly made of the terms work, cause, effect, instrument—and in the
habit of prefixing to them respectively the definitive article the—seems to be
implied a notion, of which the more closely it is examined, the more plainly will
the incorrectness be made apparent. This is that where the effect is considered as
one, there exists some one object, and no more than one, which with propriety can
be considered as its cause.17

Bentham's main argument against causal determination was epistemological,
and it comes in the following steps: i. in most cases, there is no single cause, but
causation works through a connected network of proper causes, enabling and
background conditions; ii. considering all these factors is often difficult or even
impossible, and most inferences from these supposed global causal factors to
alleged effects are invalid:

No effect […] can […] be assigned that is not the result of a multitude of
influencing circumstances: circumstances, some always in different ways
contributing to the production of it, others frequently operating in opposition to it:
contributing to it, viz. in the character of promoting and co-operating causes;
others operating in the opposite character of obstacles.

[...] In so far then as by the term cause nothing more is meant to be designated
than one alone of all those sets of co-operating circumstances, be the effect what it
may, the cause can be never of itself be adequate to the production of it: nor
between the quantum of the effect and the quantum of the cause can any
determinate proportion have place.

But of the case in which, in the extent given to the import attributed to the
word cause, the whole assemblage of these influencing circumstances is taken into
account and comprized, it seems questionable whether so much as a single

---

example would be found.

[...] Seldom indeed does it happen that, of the co-influencing circumstances, the collection made for this purpose is compleat: nor is it always that in such collection much as the principally influencing circumstance or circumstances are included.\(^\text{18}\)

Bentham rejected causal determination not on metaphysical grounds, but rather as a consequence of his scepticism towards the possibility that human mind can ever be able to grasp complex causal networks, especially when unobserved causes are at stake.

3. Bentham's indeterminism II: permanent motives as a fiction

Bentham's clearer rejection of causal determination can be found in his theory of motivation and action. He clearly claimed that there are no typical effects of motives, because each kind of motive is able to produce different kinds of effects, and there is no possibility to reliably predict which effect on actions a given motive will have.

This argument appears in the context of Bentham's criticism of moral sense and natural law theories. Bentham dubbed with pejorative labels—"principle of sympathy and antipathy", "principle of caprice", "the phantastic principle"—a bunch of related, but somewhat different, theories, defended by authors such as Shaftesbury, Hutcheson, Hume, Beattie, Price, Clarke, Wollaston.\(^\text{19}\) In Bentham's view, the common feature of those theories was the fact that they took subjective feelings of approval or disapproval of the beholder as grounds of moral assessment.\(^\text{20}\)

Against these theories, Bentham raised two main objections. First, he contended that taking subjective feelings as grounds of moral permissibility is arbitrary and tyrannical, and it should be substituted by an appeal to objective, or "extrinsic", and self-sustaining standards—the only plausible criterion of this kind being the principle of utility:

The mischief common to all these ways of thinking and arguing [...] is their serving as a cloak, and pretence, and aliment, to despotism: if not a despotism in practice, a despotism however in disposition: which is but too apt, when pretence and power offer, to show itself in practice. [...] The only right ground of action, that can possibly subsist, is, after all, the consideration of utility, which, if it is a right principle of action, and of approbation, in any one case, is so in every other. [...] Antipathy or resentment requires always to be regulated, to prevent its doing mischief: to be regulated by what? Always by the principle of utility. The principle of utility neither requires nor admits of any other regulator than itself.\(^\text{21}\)

Second, Bentham accused moral sense theorists to produce verdicts in part overlapping with the dictates of utilitarianism, without any principled justification

\(^{18}\) Bentham (1997, 140, 142).

\(^{19}\) See Bentham (1996, II, §§ 11-2, 14, 18, 21-7, 31).


for their departure from utility in certain cases: “It is manifest, that the dictates of this principle [scil. The principle of sympathy and antipathy] will frequently coincide with those of utility, though perhaps without intending any such thing.”

Moral sense theorists seemed to acknowledge utility as a standard of morals, Bentham pointed out. But in several occasions, they departed from utility, without any justification. In doing so, they produced suboptimistic, and hence objectionable, outcomes.

However, there is another shared feature that distinguishes moral sense theories from other kinds of views. Most moral sense theorists put forward a specific view of the object of moral assessment. According to this view, moral assessment focuses neither on single, specific actions, nor on kinds of action, but rather on the agent’s intentions and character traits—more precisely, on her motives as outcomes of those character traits. External behaviour is only to be considered because it is a sign of the internal character, intentions and motives of the agent whose conduct is to be assessed. Independently of their connection with the agent's character traits, actions have no moral significance. They are morally good or bad only because they manifest a morally good or bad character. Actions, and their consequences, in isolation from the character producing them, cannot be grounds of moral assessment. This view is apparent in the following passages from Hume's *Treatise of Human Nature*:

'Tis evident, than when we praise any actions, we regard only the motives that produce'd them, and consider the actions as signs or indications of certain principles in the mind and temper. The external performance has no merit. We must look within to find the moral quality. This we cannot do directly; and therefore fix our attention on actions, as external signs. But these actions are still consider'd as signs; and the ultimate object of our praise and approbation is the motive, that produc'd them.

Accordingly, mere conformity to a moral principle or rule is not enough to qualify behaviours as morally right. Absent the right motive, an action can lack what is needed to make it right. An act of gratitude performed out of mere conformity to social rules, without any sincere sentiment, is less praiseworthy than a spontaneous manifestation of a feeling of gratitude. Moreover, the presence of the right motive can make less blameworthy a wrong action: acting wrongly for good intentions can be excused, and failing to act rightly, but having the intention of doing it, can be ground of praise. Here's Hume statement of this view:

[...] when we require any action, or blame a person for not performing it, we always suppose, that one in that situation shou'd be influenc'd by the proper motive of that action, and we esteem it vicious in it to be regardless of it. If we find, upon

---

24 A larger treatment of these Benthamic objections to moral sense and natural law theories is in Pellegrino (2010, 154-98), on which this paper is partly based.
25 A contemporary conceptual analysis of the notion of ‘character trait’ is in Brandt (1992, ch. 4); see also Kupperman (1991, part I).
enquiry, that the virtuous motive was still powerful over his breast, tho' check'd in
its operation by some circumstances unknown to us, we retract our blame, and
have the same esteem for him, as if he had actually perform'd the action, which we
require of him.

It appears, therefore, that all virtuous actions derive their merit only from virtuous
motives, and are consider'd merely as signs of those motives.27

This focus on the motives of action, rather than on the action itself, or its
consequences, is typical of virtue ethics. Indeed, in some accounts, this view is
taken as the core claim of virtue ethics.28 In his criticism of moral sense theories
Bentham challenged this view, too. His argument was as follows. Actions cannot
be reliable signs of the agent's character and intentions, because there are no solid
grounds to draw a general nomological relation between (kinds of) motives and
(kinds of) actions from the connection holding between a given motive and the
ensuing action in any single case. The effects of a single motive in a specific
situation is not an adequate ground to infer the typical or general tendencies of that
kind of motive, because it appears that any kind of motive can give rise to whatever
kind of action, good or bad. This argument is presented, for instance, in the
following passages:

[…] from one and the same motive, and from every kind of motive, may proceed
actions that are good, others that are bad, and others that are indifferent. […].
From one and the same motive […] may result intentions of every sort of
complexion whatsoever. […] There is no sort of motive but may give birth to any
sort of action. […] If any sort of motive […] is either good or bad on the score of
its effects, this is the case only on individual occasions, and with individual
motives […].29

Bentham's skepticism toward the connection between motives and actions relies
on two further arguments. First, Bentham rejected an assumption of Hume's
position, i.e. the idea of a prevailing or ruling motive, able to produce a specific
kind of action in any given occasion. For Bentham, any action is the outcome of
many, even contrasting, motives, whose combined and balanced operation drives
agents to act:

When a man has it in contemplation to engage in any action, he is frequently acted
upon at the same time by the force of divers motives: one motive, or set of
motives, acting in one direction; another motive, or set of motives, acting as it
were in an opposite direction. The motives on one side disposing him to engage in
the action: those on the other, disposing him not to engage in it.30

On every such occasion, be it what it may, the action is, of course, the result of
that one motive, or that groupe of simultaneously operating motives, of which, on

27 Hume (2000, III, ii, 1, § 3-4, 307). See also Hume (2000, III, iii, 1, § 19, 373). On this
view in Hume, see Watson (1990).
28 See Watson (1990). On virtue ethics, see Annas (2006), Hursthouse (1999), Nussbaum
that same occasion, the force of influence happen to be the strongest.\textsuperscript{31}

Second, Bentham believed that very often establishing which motives to single out as the cause of a given action is quite a social matter. Motives are ascribed to people not on the ground of an allegedly impartial interpretation of their actions, but rather out of social and personal biases, and often more socially respectable motives are mentioned in place of the genuine ones:

\[
\text{[\ldots] the sort of motive to which a man's conduct will be apt to be ascribed in preference, will vary with the relative position of him to whom, on the occasion in question, it happens to speak or think of it. The best motive, that will be recognised as capable of producing the effect in question, is the motive, to which the man himself,—and, in proportion as their dispositions towards him are amicable, other men in general,—will be disposed to ascribe his conduct \[\ldots\].}
\]

Things being in this state, if, among the causes by which the conduct in question was actually produced, a motive, of a complexion sufficiently respected to be found, this is the motive, to which \[\ldots\] the conduct will be ascribed. But, if no sufficiently respected motive can be found, then, instead of the action motive, some such other motive will be looked out for and employed as being sufficiently favourable, shall by the nearness of its connection with the actual one, have been rendered most difficultly distinguishable from it. To speak shortly, if the actual motive does not come up to the purpose, another will, in the account given of the matter, be substituted to it; or, more shortly still, the motive will be changed \[\ldots\].\textsuperscript{32}

Often, agents hide their real motives declaring socially recommendable motives as the springs of their action, exaggerating the goodness of one's own motives and undervaluing the moral praiseworthiness of others' motives, and producing an ideologically loaded nomenclature of motives:

\[
\text{The bulk of mankind, ever ready to depreciate the character of their neighbours, in order, indirectly, to exalt their own, will take occasion to refer a motive to the class of bad ones as often as they can find one still better, to which the act might have owed its birth. Conscious that his own motives are not of the best class, or persuaded that if they be, they will not be referred to that class by others; afraid of being taken for a dupe, and anxious to show the reach of his penetration; each man takes care, in the first place, to impute the conduct of every other man to the least laudable of the motives that can account for it: in the next place, when he has gone as far that that way as he can, and cannot drive down the individual motive to any lower class, he changes his battery, and attacks the very class itself. To the love of reputation he will accordingly give a bad name upon every occasion, calling it ostentation, vanity, or vain-glory.}
\]

\[
\text{Partly to the same spirit of detraction \[\ldots\], may, perhaps, be imputed the great abundance of bad names of motives, in comparison of such as are good or neutral: and in particular, the total want of neutral names for the motives of sexual desire, physical desire in general, and pecuniary interest.\textsuperscript{33}}
\]

These mechanisms of social hypocrisy, Bentham concluded, made inherently

\textsuperscript{31} Bentham (1983b, 112).
\textsuperscript{32} Bentham (1983b, 112-3).
\textsuperscript{33} Bentham (1996, XI, § 17 n. g, 129-30).
problematic and doomed to failure the very inference on which Hume's view was based. Bentham argued that no external signs, emotional or physiological, can be employed as conclusive signs of the internal motives of a person:

The emotions of the body are received, and with reason, as probable indications of the temperature of the mind. But they are far enough from conclusive. A man may exhibit, for instance, the exterior appearances of grief, without really grieving at all, or at least in any thing near the proportion in which he appears to grieve. [...] Many men can command the external appearances of sensibility with very little real feeling. [...] Indications rather less equivocal may, perhaps, be afforded by the pulse. A man has not the motions of his heart at command as he has those of the muscles of his face. But the particular significance of these indications is still very uncertain. All they can express is, that the man is affected; they cannot express in what manner, nor from what cause. To an affection resulting in reality from such or such a cause, he may give an artificial bias, and represent it as if directed to such or such another object. Tears of rage he may attribute to contrition. The concern he feels at the thoughts of a punishment that awaits him, he may impute to a sympathetic concern for the mischief produced by his offence.  

Bentham pushed his scepticism till the point of denying the very possibility of introspection, till the point of claiming that our motives can be obscure even from the first person standpoint, and that we can find ourselves acting without knowing why:

But how, it may be asked, is it possible that the motive by which a man is actuated can be secret to himself? Nothing, actually, is easier; nothing is more frequent. Indeed the rare case is, not that of a man's not knowing, but that of his knowing it. It is the same with the anatomy of the human mind as with the anatomy and physiology of the human body: the rare case is, not that of a man's being unconversant with it, but that of being conversant with it. The physiology of the body is not without its difficulties; but in comparison with those by which the knowledge of the physiology of the mind has been obstructed, they are slight indeed.

Notice that Bentham's objection is not simply an instance of a sceptical view of generalization. Hume would have been ready to allow that actions are signs of the character only *ceteris paribus* in most of the cases. What Bentham is saying is not that there are exceptional cases, where certain motives fail to produce their typical effects. He is advancing a more extreme claim. He is contending that there are no typical effects of motives. Each and every motive can produce each and every action. Hume's cautious inductivism allowed that the instances in which there is a stable connection between a given kind of motive and a given kind of action are the majority of the observed cases, even though exceptions are possible. By contrast,  

---

34 Bentham (1996, VI, § 33 and n. p, 63).  
35 Bentham (1952, 235).  
36 Indeed, Hume seemed to be aware of the difficulties in introspection that Bentham emphasized, as well as of the social determination of declared motives, at least in the following passage: "Our predominant motive or intention is, indeed, frequently concealed from ourselves when it is mingled and confounded with other motives which the mind, from vanity or self-conceit, is desirous of supposing more prevalent." (Hume
Bentham’s radical scepticism led him to conclude that stable connections between certain motives and certain actions are only minor exceptions, and in the majority of cases the fact that a given motive prompted a specific agent to a given behaviour is not an instance of a general structure, but rather a particular event, to be explained only by appeal to specific and contextual circumstances. Bentham concluded that it is impossible “to know what motives have concurred or in what proportion” to the performance of a given action.\textsuperscript{37}

As a consequence, Bentham rejected the claim that long-standing or permanent kinds of motives are in operation, and that general explanations of the human conduct can be framed by appeal to them. He claimed that the very notion of a \textit{disposition}—namely the notion of “what there is supposed to be \textit{permanent} in a man’s frame of mind, where, on such or such an occasion, he has been influenced by such or such a motive, to engage in an act”—is merely the name of “a kind of fictitious entity.”\textsuperscript{38} It is a merely theoretical notion, Bentham explained, because all that appears is one single action, attended with one single train of circumstances: but from that degree of consistency and uniformity which experience has shown to be observable of in the different actions of the same person, the probable existence (past or future) of a number of acts of a similar nature, is naturally and justly inferred from the observation of one single one. Under such circumstances, such as the motive proves to be in one instance, such is the disposition to be presumed to be in others.\textsuperscript{39}

Any ascription of dispositions to individuals is the outcome of a generalization, based on the assumption that the same kind of action derives from the same kind of motive. However, as Bentham has explained in the chapter on motives in the \textit{Introduction}, this assumption is defeasible, and there is no fact of the matter on which it can be grounded. Facts concern only single actions and single motives. Accordingly, generalizations produce only “names of fictitious entities”, i.e. theoretical notions, whose validity can always be challenged by contrary evidences. This sceptical view of the link between motives and actions amounts to rejecting causal determination in the realm of action, and thereby substantiates Bentham’s indeterminism.

\textbf{4. Bentham’s indeterminism III: subjective sensibility to pleasures and pains}

The root of Bentham’s indeterministic view of the link between motives and actions lies in his conception of the nature of motives, and generally in his philosophy of mind. Despite his endorsement of the basic elements of Hume’s view of the mind and his usage of the language of the traditional faculty psychology, Bentham rejected two common assumptions of moral sense theorists, namely their appeal to complex mental items such as sentiments and durable dispositions, and their motivational pluralism (i.e. the view that many different kinds of mental items are able to motivate).

\begin{footnotesize}
\begin{itemize}
  \item \textsuperscript{37} Bentham (1983b, 15).
  \item \textsuperscript{38} Bentham (1996, 11, § 1, 125).
  \item \textsuperscript{39} Bentham (1996, 11, § 4, 126).
\end{itemize}
\end{footnotesize}
Bentham distinguished mental states along two lines. First, there are mental states directly connected with the external world, and produced by the interaction between external objects and the senses, and there are mental states obtained by the former, through inertial permanence in the mind of the original perceptions and elaboration on behalf of the mind itself. Here, Bentham employed Hume's labels, and called these two kinds of mental states impressions and ideas. Second, Bentham distinguished perceptions according to their hedonic tone: some perceptions feel as pleasant or painful, whereas some perceptions are hedonically neutral. The former are, in Bentham's language, “pathematic”, the latter “apathematic.”

Bentham denied that our mind is inhabited by more complex elements, such as sentiments, dispositions, attitudes and so on. The notions of the traditional moral psychology, he claimed, are without genuine referents, they are simply names of fictitious entities, to be paraphrased and reduced to more elementary components such as pleasant or painful, or indifferent, impressions and ideas. According to Bentham, no further analysis of mental contents is necessary once one has pointed out the role of impressions and ideas, and of pleasures and pains, because the notions of traditional psychology can be reductively paraphrased in terms of pleasures and pains. Bentham made clear the superfluity of such an analysis in the Preface of the Introduction:

As an introduction to the principles of morals, in addition to the analysis it contains of the extensive ideas signified by the terms pleasure, pain, motive, and disposition, it ought to have given a similar analysis of the not less extensive, though much less determinate, ideas annexed to the terms emotion, passion, appetite, virtue, vice, and some others, including the names of the particular virtues and vices. But as the true, and, if he conceives right, the only true groundwork for the development of the latter set of terms, has been laid by the explanation of the former, the completion of such a dictionary, so to style it, would, in comparison of the commencement, be little more than a mechanical operation.

On this view of the mind Bentham based his motivational hedonism, according to which only pleasant or painful perceptions can prompt individuals to act, and motives for action fully coincide with pleasant or painful thoughts. It is easy to allow that, if there are no other active mental states besides pleasures and pains, the latter should be responsible for action.

As previously announced, Bentham's view of pleasures and pains was a premise for his scepticism concerning any general established connection between kinds of motives and kinds of effects. On the nature of pleasures and pains, Bentham put forward three specific claims:

1. list of the sources of pleasures and pains There are typical (sources of) pleasures and pains, and Bentham gave a list of them both in the Introduction and

41 See Bentham (1997, 172).
ii. **hedonic measurement and felicific calculus** Pleasant and painful sensations may be assessed along quantitative dimensions, such as their felt intensity, their diffusion (in Bentham's language, the “extent”), their duration, and so on. Bentham wrote many pages on the quantitative assessment of pleasures and pains, and on the calculations made thereby possible.\(^{43}\)

Bentham clearly suggested that motivational hedonism is to be employed by legislators in order to provide a utilitarian code of laws with effective enforcement. Furthermore, Bentham explicitly invited legislators to ground their provisions, and the sanctions connected to them, on a view of the most diffused and relevant sources of pleasures and pains, as well as on the calculation of the effect that a given sanction can have on the behaviour of prospective misdoers:

> […] the happiness of the individuals, of whom a community is composed […] is the end and the sole end which the legislator ought to have in view: the sole standard, in conformity to which each individual ought, as far as depends upon the legislator, to be made to fashion his behaviour. But whether it be this or any thing else that is to be done, there is nothing by which a man can ultimately be made to do it, but either pain or pleasure. […]

There are four distinguishable sources from which pleasure and pain are in use to flow: considered separately, they may be termed the **physical**, the **political**, the **moral**, and the **religious**: and inasmuch as the pleasures and pains belonging to each of them are capable of giving binding force to any law or rule of conduct, they may all of them be termed **sanctions**.

> […] Pleasures and pains are the instruments [the legislator] has to work with: it behoves him therefore to understand their force, which is again, in another point of view, their value.

> […] the subject of motives is one with which it is necessary to be acquainted, in order to pass a judgment on any means that may be proposed for combating offences […]\(^{44}\)

Bentham's insistence on the relevance of protecting settled expectations and security through law is obviously related to the importance of both pleasures of expectations and pains of disappointment. Likewise, Bentham's claim that legal sanctions should be strictly proportional obviously relies on his confidence that it is possible to measure up pleasures and pains, as the two elements to be made proportional are the expected pleasures for the misdoer ensuing from the possible offense and the expected pains promised to potential misdoers by the legislator. A sanction is proportional when the threatened pain is able to outweigh the potential pleasure deriving from the commission of an offense, thereby preventing potential offenders.

\(^{43}\) See Bentham (1996, 4, 38-41).

This connection between law enforcement and motivational hedonism might appear to give support to a deterministic view of human action. In suggesting the legislators to rely on a list of the typical sources of pleasures and pains and on a measurement of the impact of expected pleasures and pains, Bentham implicitly intimates that the same kind of sanction can prevent each and every individual from committing evil. This amounts to claiming that the same kind of motive—the desire to avoid certain pains, the pains of typical sanctions (such as fines, imprisonment, and so on)—always produces the same kind of effect, i.e. abstention from evil-doing.

However, Bentham presented a third claim concerning pleasures and pains, that substantially weakens the determinist consequences of the first two claims:

iii. *hedonic subjective sensibility* The capacity to feel the intensity of certain pleasures or pains, and to be moved by them, is subjectively determined, namely it varies depending on many circumstances.

This view is clearly presented in the following passage:

Pain and pleasure are produced in men's minds by the action of certain causes. But the quantity of pleasure and pain runs not uniformly in proportion to the cause; in other words, to the quantity of force exerted by such a cause. [...] But in the same mind such and such causes of pain or pleasure will produce more pain or pleasure than such or such other causes of pain or pleasure: and this proportion will in different minds be different. [...] One man, for instance, may be most affected by the pleasures of the taste; another by those of the ear. So also, if there be a difference in the nature or proportion of two pains or pleasures which they respectively experience from the same cause [...]. From the same injury, for instance, one man may feel the same quantity of grief and resentment together as another man: but one of them shall feel a greater share of grief than of resentment: the other, a greater share of resentment than of grief. 45

The subjective sensibility to pleasures and pains has many relevant consequences. First of all, it alters the meaning of Bentham's list of the typical sources of pleasures and pains. As clearly stated in the passage above, certain individuals will be less sensible to certain pleasures or pains, and what for certain subjects is most pleasant may be less pleasant, or even painful, for others. Accordingly, Bentham's list of pleasures and pains in chapter V of the *Introduction* is to be considered simply a general assemblage of whatever *might* be pleasant or painful, and not of what *will surely be* pleasant or painful. The usage of this list as a basis for assigning sanctions to laws, then, is much more limited than one could have expected at first sight. It might be conceivable that a particularly tough person, with a slow sensibility to suffering, may be much less touched by the threat of imprisonment than a more sensible, and delicate, individual. The same sanction aimed at punishing the same misconduct—for instance a case of bribery—can be overwhelmingly frightening for the an individual, and indifferent to another. Accordingly, the legislator cannot be sure that the same sanction will work with the same degree of effectivity in any case.

Also, subjective sensibility makes hedonic calculations possible only from a subjective perspective. Any single individual can assess the intensity of each of her own pleasures and pains, and she can compare pleasures and pains, establishing which pleasure overrides a pain to be suffered for the sake of obtaining it, and vice versa. However, this sort of comparison between different pleasures and pains cannot be done from an external perspective, i.e. from an objective point of view, or at least it cannot be done with the same accuracy, because it might be the case that the same pleasure or pain has different intensities when felt by different individuals.

Bentham's insistence on subjective hedonic sensibilities, then, makes the reliance of legislation on hedonic psychology defeasible. However, Bentham did never give up on his idea that sanctions should be conceived in terms of pleasures and pains, and that the legislator should employ a hedonistic motivational psychology in order to enforce law. Still, he pointed out that the legislator should act upon a very general knowledge, and legislative interference should be had only in very general cases, abstaining from any interference in specific and particular situations:

It is a standing topic of complaint that a man knows too little of himself. Be it so: but is it so certain that the legislator must know more? It is plain, that of individuals the legislator can know nothing: concerning those points of conduct which depend upon the particular circumstances of each individual, it is plain, therefore, that he can determine nothing to advantage. It is only with respect to those broad lines of conduct in which all persons, or very large and permanent descriptions of persons, may be in a way to engage, that he can have any pretence for interfering; and even here the propriety of his interference will, in most instances, lie very open to dispute. At any rate, he must never expect to produce a perfect compliance by the sanction of which he is himself the author. [...] With what chance of success, for example, would a legislator go about to extirpate drunkenness and fornication, by dint of legal punishment? Not all the tortures that ingenuity could invent would compass it: and, before he had made any progress worth regarding, such a mass of evil would be produced by the punishment, as would exceed, a thousand-fold, the utmost possible mischief of the offence.46

Here the liberal consequences of Bentham's indeterminism should be clear. The wavering of the subjective sensibilities to the pleasures and pains makes the reaction of agents to sanction unpredictable, and this, in its turn, makes ineffective, and even counterproductive, sanction in certain areas. The sphere where Bentham suggested that the knowledge of the legislator is too small and indeterminate to produce effective restraint is the realm of private ethics. 47 Bentham was clear in saying that in those areas utilitarianism still holds as a normative standard, but legislative interference is ineffective:

There is no case in which a private man ought not to direct his own conduct to the production of his own happiness [...]: but there are cases in which the legislator ought not (in a direct way at least, and by mean of punishment applied

---

47 On the distinction between private ethics and the art of government or of legislation, see Bentham (1996, 17, §§ 2-4, 282-3).
immediately to particular individual acts) to attempt to direct the conduct of the several other members of the community.\footnote{48 Bentham (1996, 17, § 8, 285).}

Here, Bentham clearly stated a version of the liberal distinction between private and public sphere. His indeterministic view of action, and the unpredictability that this view entails, is the main premise of this distinction.

5. Bentham's non-manipulative liberalism (a better rationale for the best-judge principle)

In the passage above, Bentham set limits to legislative interference on the basis of the fact that the legislator lacks the knowledge necessary to produce good effects through sanctions. A similar view is the best judge principle—another traditionally liberal principle—according to which individuals have better knowledge of their interest than any other person, who, for this reason, should abstain from any paternalistic interference. Applied to the realm of politics, the principle dictates that no state interference on private spheres it is legitimate, except in matters of basic justice with the aim of regulating coexistence among free and equal citizens.\footnote{49 See Caney (1991, 463-5), Goodin (1990), Guest (1989).} Bentham stated this principle deducing it from his indeterministic view and his conception of the subjective sensibility to pleasures and pains:

Proportioned to the differences between the particular sensibilities of the several persons in question will be the absurdity of him who, in a case in which the agent himself were the only person whose well-being were in question, should in the character of a Moralist or say a Deontologist, prescribe exactly the same line of conduct to be observed by every man.

With a benefit of a certain degree of experience it may be delivered in the character of a general proposition [that] every man is a better judge of what is conducive to his own well-being than any other can be.\footnote{50 Bentham (1983c, I.3, 131).}

In many later formulations, the best-judge principle seems to rest on an epistemological argument: only individuals have access to their mental states; therefore, only individuals can establish whether their interests, in the form of pleasure, happiness or satisfaction of desires, are promoted.\footnote{51 See Goodin (1990, 183).} This argument has been criticized in a number of ways. First, interests and mental states, though connected, are different things, and it is conceivable that even though mental states are inherently private, promotion of interests should necessarily be a public matter, as it is related to empirical, or objective, issues. Second, it might be showed that individuals' knowledge and promotion of their interests is often affected by many biases and obstacles – such as irrational temporal preferences, irrational discounting of probabilities, akritic indolence, collective coordination problem. In this sort of cases, public knowledge of individual interests, and the ensuing promotion, would be better than individual action.\footnote{52 See Goodin (1990, 189-91).}
Bentham's indeterministic view of the connection between motives and actions provides a better rationale for the best-judge principle. Bentham showed that even though individual interests could be better promoted through governmental interference, in many cases there is no way to do it—or at least, there is no way to promote individual interests better than how individuals would do. In other words, Bentham provided a new epistemological reason in favour of the best-judge principle, a reason grounded on the supposed unpredictability of individual behaviour from an external perspective.

In a sense, Bentham seemed also to be aware of the fact that the best-judge principle has only a limited validity. He was clear that there are both cases in which individuals would better promote their interests and cases in which state intervention would do better than private action. What is important is that Bentham provided a principled way to demarcate those cases, and to justify state interference or private action. State activism is needed when individual actions are predictable enough to be affected by general sanctions; legislative interference is ineffectual, and obnoxious, when there is no way to promote reliably individual interests with the means of legislation.

Bentham's scepticism towards the effectiveness of legal sanctions in many cases does not weaken his commitment to the project of a utilitarian legislation. Indeed, Bentham's view of the limits of the predictability of human action explains some seemingly puzzling claims of his. As many recent interpreters have emphasized, Bentham did not maintain that the legislator should directly promote the greatest happiness of the greatest number. Rather, he suggested that the legislator should aim at promoting four subordinate sub-ends, namely subsistence, abundance, equality and security. Promotion of those sub-ends should be consistent, and it seems that Bentham believed that security and subsistence have priority over equality, and equality over abundance.

This claim might be considered inconsistent with Bentham's insistence on the relevance of the principle of utility. It might be asked: Why Benthamic legislator should abstain from direct promotion of the greatest happiness of the greatest number? It might be argued that there are cases in which none of the four sub-ends would promote the greatest happiness of the greatest number. It is easy to imagine, for instance, a case where greater inequality would maximize happiness. In many cases, also, a sacrifice of security and subsistence for some people may serve to the maximization of happiness for the majority.

To these objections, sometimes Bentham seemed to have provided responses in terms of psychological premises. For instance, he claimed that no detraction to security and subsistence can be compensated by increased happiness—due to the fundamental role of security of expectations and subsistence in the psychological dynamics of the average human beings. Also, Bentham suggested that inequalities can create more pain than pleasure. His indeterministic view of human action can be considered a further premise used to justify the indirect route that Bentham established for the utilitarian legislators. Subsistence, security, equality and abundance might be the only motives whose connection with action is established.
enough to guarantee that legislative interference is both useful and effective. In other words, when the legislator threatens to diminish the security, subsistence, equality and abundance of the offenders, her sanctions are effective in any case. Moreover, when she makes her threats in order to promote more security, subsistence, equality and abundance for the citizens, her interference is justified because it promotes the greatest happiness of the greatest numbers.

Bentham's nuanced approach produces a principled way to both limit and legitimate state interference, a way that is alternative to more known liberal strategies. Bentham did not distinguish the right from the good, nor he had a theory about the neutrality of the state. Instead, he justified the limit and the scope of legislative and state interference by invoking psychological premises about the connection between motives and action, and by establishing a minimal list of primary goods that legislation should promote.

Importantly, Bentham's indeterministic view of the connection between motives and actions makes manipulation impossible. Bentham would have been sceptical on the very possibility that a happiness pill could work, for the simple reason that there would be no sure grounds to predict its effects. People react differently to pleasures and pains, and for this reason they react differently to sanctions. Arguably, they would react differently also to chemical manipulation of their brains, and these different reactions would make manipulation ineffective in many cases. Independently of its accuracy in the light of the results of contemporary physiology, which is not at stake here, Bentham's indeterministic approach makes manipulative illiberalism impossible. His view of motives and their connection with actions makes impossible the nightmare of an illiberal mass treatment with medicines: if there are no stable connections between motives and actions, then it is impossible for any medicine having uniform effects over persons.

References


